Certificate of facsimile transmission for Documents for Amendment/Responses 9/05 and 11/05 for application 10/037,718 November 21, 2005. Submitted by fax to USPTO at 571-273-8300

Certificate of Facsimile Transmission

The following documents were submitted by facsimile today, November 21, 2005 by me, Robert O. McGinnis to the USPTO. These documents were faxed to 571-273-8300 by me.

The documents are as follows:

1) The Risch paper (Risch and Merikangas, Future of Genetic Studies of Complex Diseases, Science vol 273, pp. 1516 and 1517; 2 pages total)

2) The Chee paper (Chee, et al, Accessing Genetic Information with High-Density DNA Arrays, Science vol 274, page 610 and 613, 2 pages total)

3) The Krugkyak paper (Kruglyak, L., The use of a genetic map of bi-allelic markers in linkage studies, Nature Genetics, vol 17, pp. 21-24, 4 pages total).

4) The Saiki paper (Saiki, et al, Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes, Proc. Natl. Acad. Sci. USA, vol 86, p 6230, 1 page total).

5) Excerpt from page 1228 from Merriam-Webster's Collegiate Dictionary (TENTH EDITION) of definition of **thousand, thousands** (1 page total).

6) Illustration entitled Illustration of an Example Set/Subset N-covering using a CL-F map (1 page total).

7) Accompanying letter (signed) 2 pages

Total Sheets (or pages), including this Certificate of Facsimile Transmission, 14 sheets.

R O. McGinnis

November 21, 2005
Robert O. McGinnis
Agent of Record
Reg. No. 44, 232
Ph. 406-522-9355

Copies Of Documents To Be Used By Examiner In Previously Filed Amendment/Responses for application No. 10/037, 718 1 of 2

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of:
Ralph Evan McGinnis and Robert Owen McGinnis
Application No.: 10/037,718
Filed: 01/04/2002

Title of the Invention: TWO-DIMENSIONAL LINKAGE STUDY METHODS AND RELATED INVENTIONS

Art Unit 1637

Examiner: Horlick, K.

Honorable Commissioner for Patents
submitted by fax to 571-273-8300

COPIES OF DOCUMENTS TO BE USED BY EXAMINER IN PREVIOUSLY FILED AMENDMENT/RESPONSES

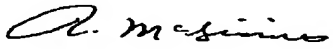
Sir:
Enclosed herewith are copies of pages of published papers, an excerpt from a dictionary and an illustration to be used by the Examiner in connection with two recently filed Amendment/Responses. The Amendment/Responses are an Amendment/Response, RCE filed on September 13, 2005 and a Supplemental Amendment/Response filed in November 2005. The applicants respectfully request that these copies be forwarded to the Examiner. Total of 11 pages of copies of documents.

The copies are as follows:
1) The Risch paper (Risch and Merikangas, Future of Genetic Studies of Complex Diseases, Science vol 273, pp. 1516 and 1517; 2 pages total)
2) The Chee paper (Chee, et al, Accessing Genetic Information with High-Density DNA Arrays, Science vol 274, page 610 and 613, 2 pages total)
3) The Krugkyak paper (Kruglyak, L., The use of a genetic map of bi-allelic markers in linkage studies, Nature Genetics, vol 17, pp. 21-24, 4 pages total).
4) The Saiki paper (Saiki, et al, Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes, Proc. Natl. Acad. Sci. USA, vol 86, p 6230, 1 page total).

Copies Of Documents To Be Used By Examiner In Previously Filed Amendment/Responses for application No. 10/037, 718 2 of 2

5) Excerpt from page 1228 from Merriam-Webster's Collegiate Dictionary (TENTH EDITION) of definition of **thousand, thousands** (1 page total).
6) Illustration entitled Illustration of an Example Set/Subset N-covering using a CL-F map (1 page total).

Respectfully submitted,



Robert O. McGinnis, Reg. No. 44, 232
Ph. 406-522-9355
November 21, 2005

Merriam-Webster's Collegiate Dictionary

TENTH EDITION

Excerpt from page 1228

thought–out \-'aůt\ *adj* (1870) : produced or arrived at through mental effort and esp. through careful and thorough consideration
thought·way \-ˌwā\ *n* (ca. 1944) : a way of thinking that is characteristic of a particular group, time, or culture
thou·sand \'thaů-zᵊn(d)\ *n, pl* **thousands** *or* **thousand** [ME, fr. OE *thūsend;* akin to OHG *dūsunt* thousand, Lith *tūkstantis,* and prob. to Skt *tavas* strong, L *tumēre* to swell — more at THUMB] (bef. 12c) **1** — see NUMBER table **2** : a very large number ⟨~s of ants⟩ — **thousand** *adj* — **thou·sand·fold** \-zᵊn(d)-ˌfōld\ *adj or adv* — **thou·sandth** \-zᵊn(t)th\ *adj or n*
Thousand Island dressing *n* [*Thousand Islands,* islands in the St. Lawrence River] (1924) : mayonnaise with chili sauce and seasonings (as chopped pimientos, green peppers, and onion)
thousand–leg·ger \ˌthaů-zᵊn(d)-'le-gər, -'lā-\ *n* (1914) : MILLIPEDE
thousands place *n* (1937) : the place four to the left of the decimal point in a number expressed in the Arabic system of writing numbers
Thra·cian \'thrā-shən\ *n* (1569) **1** : a native or inhabitant of Thrace **2** : the Indo-European language of the ancient Thracians — see INDO-EUROPEAN LANGUAGES table — **Thracian** *adj*
[1]thrall \'thrȯl\ *n* [ME *thral,* fr. OE *thræl,* fr. ON *thræll*] (bef. 12c) **1 a** : a servant slave : BONDMAN; *also* : SERF **b** : a person in moral or mental servitude **2 a** : a state of servitude or submission ⟨in ~ to his emotions⟩ **b** : a state of complete absorption ⟨mountains could hold me in ~ with a subtle attraction of their own —Elyne Mitchell⟩ — **thrall** *adj* — **thrall·dom** *or* **thral·dom** \'thrȯl-dəm\ *n*
[2]thrall *vt* (13c) *archaic* : ENTHRALL, ENSLAVE

The use of a genetic map of biallelic markers in linkage studies

Leonid Kruglyak

Improvements in genetic mapping techniques have driven recent progress in human genetics. The use of single nucleotide polymorphisms (SNPs) as biallelic genetic markers offers the promise of rapid, highly automated genotyping. As maps of SNPs and the techniques for genotyping them are being developed, it is important to consider what properties such maps must have in order for them to be useful for linkage studies. I examine how polymorphic and densely spaced biallelic markers need to be for extraction of most of the inheritance information from human pedigrees, and compare maps of biallelics with today's genome-scanning sets of microsatellite markers. I conclude that a map of 700–900 moderately polymorphic biallelic markers is equivalent—and a map of 1,500–3,000 superior—to the current 300–400 microsatellite marker sets.

The revolution in human genetics that has unfolded over the past decade and a half has been driven largely by the development of genetic maps. The original concept was proposed by Botstein *et al.*, with restriction fragment length polymorphisms (RFLPs) as markers[1]. The first human RFLP was quickly identified[2], and Huntington's disease soon became the first autosomal disorder linked to an anonymous DNA marker[3]. The first RFLP map of the human genome followed shortly[4]. RFLPs were based on a variety of polymorphisms at the sequence level (single nucleotide changes, insertions and deletions, repeat length polymorphisms) and were assayed by Southern hybridization. Although a great advance, RFLPs were often not very polymorphic, and they were costly and time-consuming to develop and assay in large numbers. Nevertheless, these markers made human molecular genetics a reality and led to the mapping of a number of important mendelian diseases.

The next major advance came with the discovery and development of microsatellites (STRs or SSLPs) as markers[5]. These loci are abundant, have fairly high polymorphism rates and can be assayed by PCR, leading to lower cost and a greater degree of automation. Dense maps of microsatellites are now available[6,7], allowing simple mendelian diseases to be mapped with relative ease and enabling first searches for genetic causes of complex diseases by genome scan. However, the requirements to assay the loci on gels and to distinguish several length-based alleles make it hard to fully automate the genotyping process, and typing large numbers of individuals for markers covering the genome remains beyond the resources of all but a few labs. There is thus a need to move beyond this current technology.

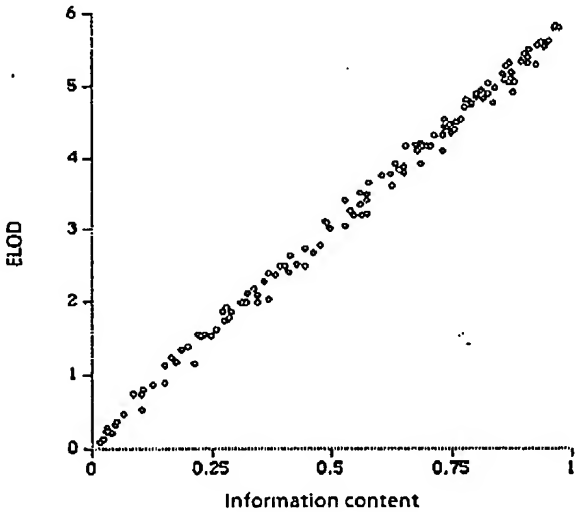


Fig. 1 Expected lod score (ELOD) for a dominant locus is plotted against information content. Each circle represents the results of a simulation for one of 130 maps, as described in Methods. The solid line shows the expected linear correlation if information content of 0 corresponds to an ELOD of 0 and information content of 1 corresponds to the maximum achievable ELOD of 6.02 in these pedigrees.

Recent attention has focused on the use of single nucleotide polymorphisms (SNPs) as genetic markers. At first glance, this may appear to represent a step back to the days of low polymorphism rates characteristic of RFLPs. However, modern technology should allow efficient assays of SNPs in numbers sufficiently large to offset their lower polymorphism rates, as discussed below. SNPs offer a number of important advantages over microsatellites. They are highly abundant, with classic estimates of more than 1 per 1,000 base pairs, or more than 3 million in the genome[8,9]. To date, more than 1,000 PCR-amplifiable SNP markers have been discovered and mapped (D. Wang, pers. comm.). Because SNPs have only two (common) alleles (hence the term 'biallelics'), genotyping them requires only a plus/minus assay rather than a length measurement, permitting easier automation. Several non-gel-based assays have been proposed[10–14], with high-density oligonucleotide arrays currently showing great promise for typing large numbers of biallelic markers in parallel[15,16].

Here I consider the feasibility of carrying out linkage studies with a genetic map based on biallelic markers. The key questions are: What level of polymorphism is required? and How many markers adequately cover the genome? These questions are addressed below.

Assumptions

The effects of marker density and polymorphism were examined by simulating pedigree genotype data and measuring the information content[17,18] for a broad range of map densities and polymorphism levels (see Methods for simulation details). Information content measures the fraction of inheritance information extracted by the map relative to that which

Whitehead Institute for Biomedical Research, 1 Kendall Square, Bldg. 300, Cambridge, Massachusetts 02139, USA. email: leonid@genome.wi.mit.edu

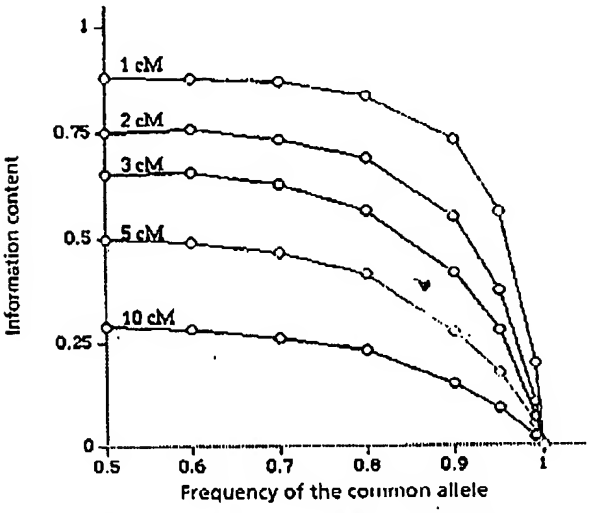


Fig. 2 Information content for five map densities is plotted against the frequency of the more common of the two alleles of a biallelic marker. The circles show actual simulation data points.

would be extracted by an infinitely dense polymorphic map. Thus, an information content of 1 reflects complete information, whereas an information content of 0 reflects no information. Information content incorporates both marker density and polymorphism in a single general measure of map quality that is independent of assumptions about a particular disease locus. It also closely predicts the power of a map to detect linkage—for example, as measured by the expected lod score (ELOD; Fig. 1).

The markers were assumed to be evenly spaced, and information content was measured at a location halfway between two markers, where it is expected to be lowest. For clarity, a single pedigree structure is used throughout: first-cousin pairs with parents but not grandparents available for genotyping. Extensive simulations show that although the absolute numbers differ somewhat for other pedigree structures, all the main conclusions about the relative importance of marker polymorphism and density continue to hold.

How polymorphic do biallelic markers need to be?

Biallelic markers vary in their rates of polymorphism: the more common allele can range in frequency from 50% to nearly 100%. In considering a map of biallelic markers, it is important to ask whether only near-perfect (50–50) biallelics are useful or whether less polymorphic markers can provide comparable amounts of information. To answer this question, I measured information content in simulations of maps of biallelic markers with varying degrees of polymorphism.

Table 1 • Information content for biallelics

	allele distribution				
spacing (cM)	50-50	60-40	70-30	80-20	90-10
1	0.88	0.88	0.87	0.84	0.73
2	0.75	0.76	0.73	0.69	0.55
3	0.65	0.65	0.63	0.56	0.42
4	0.58	0.56	0.53	0.48	0.34
5	0.50	0.49	0.46	0.41	0.27
6	0.45	0.43	0.41	0.36	0.24
7	0.39	0.39	0.37	0.32	0.22
8	0.35	0.35	0.33	0.28	0.19
9	0.32	0.31	0.29	0.25	0.17
10	0.29	0.28	0.26	0.23	0.15

The results (Fig. 2, Table 1) clearly indicate that at higher map densities, allele frequency has only a small effect on information content in the range of frequency distributions from 50-50 to 80-20. Specifically, a 1-cM map of 60-40 biallelics provides an information content of 0.88, essentially the same as perfect 50-50 biallelics at this density, while 70-30 biallelics provide an information content of 0.87, and 80-20 biallelics provide an information content of 0.84. The information content drops to 0.73 for 90-10 biallelics. Thus, the use of biallelic markers with frequency distribution as skewed as 80-20 leads to little reduction in the information content of a dense map. For sparser maps of 5–10 cM, a similar conclusion holds for marker allele frequency distributions as skewed as 70-30.

How dense does a map of biallelic markers need to be?

Although there is a limit on how polymorphic a biallelic marker can be (a 50-50 distribution of the two alleles), there is essentially no theoretical limit on map density (or marker number), as reasonably polymorphic SNPs can be found roughly every 1 kb, or about 3 million times in the human genome (see above). Thus, one answer to how many markers are needed is that more is always better[1]. For common linkage study designs, however, the addition of markers provides diminishing returns once most of the inheritance information has been extracted. As shown above, a 1-cM

Table 2 • Information content for microsatellites

	allele number			
spacing (cM)	3	4	5	10
1	0.93	0.95	0.96	0.97
2	0.87	0.90	0.91	0.94
3	0.80	0.84	0.87	0.90
4	0.74	0.78	0.81	0.87
5	0.68	0.75	0.78	0.82
6	0.64	0.70	0.73	0.80
7	0.58	0.64	0.69	0.76
8	0.53	0.61	0.66	0.72
9	0.49	0.58	0.62	0.69
10	0.45	0.54	0.58	0.68

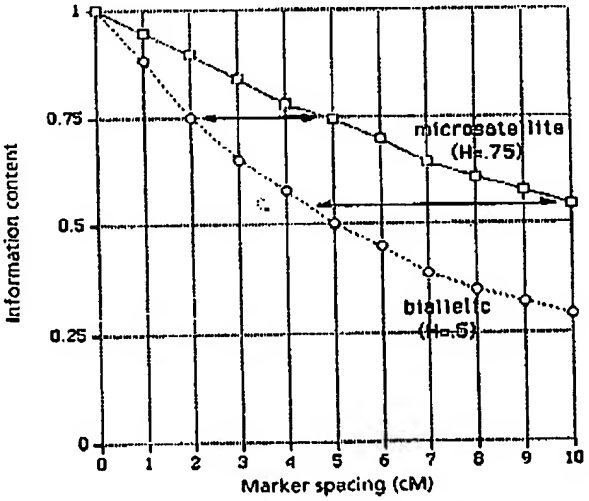


Fig. 3 Information content is plotted against marker spacing for selected microsatellite (heterozygosity H=0.75) and biallelic (heterozygosity H=0.5) markers. Arrows connect the points on the two curves where information content reaches 0.75 (top) and 0.54 (bottom), the values for 5-cM and 10-cM microsatellite maps, respectively.

map of 50–50 biallelic markers extracts 88% of the available information, and it is unlikely that higher information content is needed in an initial screen for linkage. What is the informational cost of decreasing the density of the map? Simulation results (Fig. 2) show that map density plays a more critical role than marker polymorphism. A 2-cM map provides information content of 0.75, a 3-cM map 0.65 and a 5-cM map 0.50. Together with the results of the previous section, these numbers lead to the conclusion that for initial linkage studies it is desirable to screen a dense (1–2-cM) map of moderately polymorphic (50–50 to 80–20) biallelic markers. Interesting regions can then be followed up with all available (biallelic and microsatellite) markers.

It is worth noting that there are two separate issues regarding map density: how many markers exist and how many markers can be genotyped rapidly and cost-effectively. Although current microsatellite maps cover the genome at an average spacing of less than 1 cM (with more than 5,000 markers in the final Généthon map alone[7]), genotyping more than a few hundred markers in a large collection of families remains beyond the power of today's technology and research budgets. Thus, the practical limit on the number of biallelic markers will depend on the techniques for marker development and genotyping. Nonetheless, it is interesting to compare such maps with current maps of microsatellite markers. Such a comparison is carried out in the next section.

Comparison of maps based on biallelics and microsatellites

Current genome scans typically employ a 10-cM map of microsatellite markers for the initial screen[19,20], followed by denser coverage of regions that yield interesting results. (Although one could employ a 'staged search' strategy of starting with a sparser 20–40-cM map and then increasing the density in all moderately positive regions[21,22], economies of scale in large genotyping labs usually argue for a one-stage initial scan: using a single optimized set of markers for all projects is more efficient than 'filling in' different regions for each.) Microsatellite markers typically vary between 0.65 and 0.8 in heterozygosity (for instance, an average of 0.7 in the final Généthon map[7]), and for simplicity I will use microsatellites with four equally frequent alleles (heterozygosity of 0.75) as representative in the following comparisons with biallelics with two equally frequent alleles (heterozygosity of 0.5); results for other values are given in Tables 1 and 2. Intuitively, one would expect two closely linked biallelics to provide the same information as one microsatellite, and simulations largely confirm this intuition. A 10-cM map of microsatellites achieves information content of 0.54 (Fig. 3). The same information content is provided by a 4.5-cM map of biallelic markers. A denser 5-cM microsatellite map achieves an information content of 0.75, as does a 2-cM map of biallelics. In general, maps of biallelic markers at about 2.25–2.5 times the density of microsatellites provide a comparable information content. A 10-cM map of 300 microsatellite markers can therefore be replaced by a 4-cM map of 750 biallelic markers. These conclusions are in rough agreement with the results of an earlier study of the trade-off between marker spacing and polymorphism[23].

As technology improves, it is likely that screening a much denser map of biallelic markers will be cheaper and easier than carrying out today's genome scans employing microsatellites[15,16]. There are reasons to employ such denser maps. As shown above, current scan densities lead to considerable loss of information. This problem is more serious for data-sets consisting of more distantly related affecteds or of progeny of consanguineous marriages used in homozygosity mapping[24]. It is therefore worth noting that a 1-cM map of biallelics (about 3,000 markers) yields much higher information content than a 10-cM map of microsatellites (0.88 vs. 0.54), and is superior to a 5-cM microsatellite map (0.88 vs. 0.75).

Practical linkage analysis using biallelic markers

Because of the lower polymorphism rates of biallelic markers, it is critical to consider many linked markers simultaneously; indeed, all the above results assume complete multipoint analysis of all markers on a chromosome. Such multipoint analysis is even more important for biallelics than for microsatellites. Fortunately, recently developed algorithms and software allow multipoint analysis with an essentially unlimited number of linked markers to be carried out for sib pairs[17] as well as for general pedigrees of moderate size[18]. These methods can also be used for automatic haplotype reconstruction, avoiding the tedious prospect of haplotyping many biallelics by hand. The one remaining challenge is extending multipoint analysis with many markers to large multi-generational families, although even here the situation is improving[25].

Discussion

The results presented here clearly demonstrate that the use of a genetic map of biallelic markers for linkage studies is feasible on theoretical grounds. It is not necessary to find only 'perfect' 50-50 biallelics: markers with allele frequency distributions as skewed as 70-30 or even 80-20 are almost as useful in a dense map. This result should allay the concern that markers discovered in one population may not be sufficiently informative in other populations with different allele frequencies. A 1–2-cM map of moderately polymorphic biallelic markers is superior to today's microsatellite screening sets for extracting inheritance information and should provide a more efficient tool for initial genome scans.

Even denser maps should enable novel study designs for dissecting genetically complex phenotypes. In particular, genome scans for linkage disequilibrium (LD) and association may become practical[26–28]. Because LD mapping relies on detecting recombinationally conserved regions around an ancestral mutation, the required map density will vary with the age and history of the study population, with very dense maps (spacing of 10 kb or less) likely to be needed for LD scans in a mixed general population. A more promising approach may be to screen in parallel functional (coding) biallelic polymorphisms in many genes for direct association (rather than LD) with disease[26–28].

Maps of biallelic markers and the technology to genotype them should be forthcoming[15,16], and the resulting progress in human genetics will be exciting to watch.

Methods

Simulations. Segregation of chromosomes of 100-cM length with evenly spaced markers was simulated. For biallelics, the frequencies of the common allele were 0.5, 0.6, 0.7, 0.8, 0.9, 0.95 and 0.99. For microsatellites, equally frequent alleles were assumed, with allele numbers of 3, 4, 5, 10, 20 and 100. Marker spacings of 1, 2, . . . , 10 cM were examined. Each simulation consisted of 100 replicates of 10 cousin pairs each. Information content was computed with GENEHUNTER[18]. Information content was measured halfway between the two markers closest to the middle of the chromosome. For ELOD computation, a dominant disease locus with full penetrance, no phenocopies and allele frequency of 0.001 was assumed to lie halfway between two markers, and chromosomes were simulated assuming that both cousins were affected. GENEHUNTER was used to compute multipoint lod scores. The relationship between information content and ELOD is preserved for other assumptions about the disease locus (data not shown). Simulation software used to generate the data is available from the author and can be used to explore additional map properties and pedigree structures.

Acknowledgements

I thank M. Daly, E. Lander and D. Wang for helpful discussions and comments on the manuscript. This work was supported in part by a Special Emphasis Research Career Award from NHGRI (HG00017).

1. Botstein, D., White, D.L., Skolnick, M. & Davis, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
2. Wyman, A.R. & White, R.W. A highly polymorphic locus in human DNA. *Proc. Natl. Acad. Sci. USA* **77**, 6754–6758 (1980).
3. Gusella, J.F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238 (1983).
4. Donis-Keller, H. *et al.* A genetic linkage map of the human genome. *Cell* **51**, 319–337 (1987).
5. Weber, J.L. & May, P.E. Abundant class of human DNA polymorphism which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**, 388–396 (1989).
6. Cooperative Human Linkage Center. A comprehensive human linkage map with centimorgan density. *Science* **265**, 2049–2054 (1994).
7. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5264 microsatellites. *Nature* **380**, 152–154 (1996).
8. Hofker, M.H. *et al.* The X chromosome shows less genetic variation at restriction sites than the autosomes. *Am. J. Hum. Genet.* **39**, 438–451 (1986).
9. Cooper, D.N., Smith, B.A., Cooke, H.J., Niemann, S. & Schmidtke, J. An estimate of unique DNA sequence heterozygosity in the human genome. *Hum. Genet.* **69**, 201–205 (1985).
10. Nickerson, D.A. *et al.* Automated DNA diagnostics using an ELISA-based oligonucleotide ligation assay. *Proc. Natl. Acad. Sci. USA* **87**, 8923–8927 (1990).
11. Livak, K.J., Marmaro, J. & Todd, J.A. Towards fully automated genome-wide polymorphism screening. *Nature Genet.* **9**, 341–342 (1995).
12. Saiki, R.K., Walsh, P.S., Levenson, C.H. & Erlich, H.A. Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. *Proc. Natl. Acad. Sci. USA* **86**, 6230–6234 (1989).
13. Syvanen, A.-C., Aalto-Setala, K., Harju, L., Kontula, K. & Soderlund, H. A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. *Genomics* **8**, 684–692 (1990).
14. Wu, D.Y., Ugozzoli, L., Pal, B.K. & Wallace, R.B. Allele-specific enzymatic amplification of β-globin genomic DNA for diagnosis of sickle cell anemia. *Proc. Natl. Acad. Sci. USA* **86**, 2757–2760 (1989).
15. Wang, D. *et al.* Toward a third generation genetic map of the human genome based on biallelic polymorphisms. *Am. J. Hum. Genet.* **59**, A3 (1996).
16. Chee, M. *et al.* Accessing genetic information with high-density DNA arrays. *Science* **274**, 610–614 (1996).
17. Kruglyak, L. & Lander, E.S. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am. J. Hum. Genet.* **57**, 439–454 (1995).
18. Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. & Lander, E.S. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58**, 1347–1363 (1996).
19. Reed, P.W. *et al.* Chromosome-specific microsatellite sets for fluorescence-based, semi-automated genome mapping. *Nature Genet.* **7**, 390–395 (1994).
20. Dubovsky, J., Sheffield, V.C., Duyk, G.M. & Weber, J.L. Sets of short tandem repeat polymorphisms for efficient linkage screening of the human genome. *Hum. Mol. Genet.* **4**, 449–452 (1995).
21. Elston, R.C. Designs for the global search of the human genome by linkage analysis. In *Proceedings of the 16th International Biometrics Conference* 39–51 (Hamilton, New Zealand, 1992).
22. Brown, D.L., Gorin, M.B. & Weeks, D.E. Efficient strategies for genomic searching using the affected-pedigree-member method of linkage analysis. *Am. J. Hum. Genet.* **54**, 544–552 (1994).
23. Terwilliger, J.D., Ding, Y. & Ott, J. On the relative importance of marker heterozygosity and intermarker distance in gene mapping. *Genomics* **13**, 951–956 (1992).
24. Lander, E.S. & Botstein, D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236**, 1567–1570 (1987).
25. O'Connell, J.R. & Weeks, D.E. The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genet.* **11**, 402–408 (1995).
26. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
27. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
28. Collins, F.S. Positional cloning moves from perditional to traditional. *Nature Genet.* **9**, 347–350 (1995).

Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 6230–6234, August 1989
Genetics

Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes

(polymerase chain reaction/"reverse dot blots"/nonradioactive detection/*HLA-DQA* locus/β-thalassemia)

RANDALL K. SAIKI*, P. SEAN WALSH*, COREY H. LEVENSON†, AND HENRY A. ERLICH*

Departments of *Human Genetics and †Chemistry, Cetus Corp., 1400 Fifty-Third Street, Emeryville, CA 94608

Communicated by Hamilton O. Smith, May 9, 1989 (received for review March 2, 1989)

ABSTRACT The analysis of DNA for the presence of particular mutations or polymorphisms can be readily accomplished by differential hybridization with sequence-specific oligonucleotide probes. The *in vitro* DNA amplification technique, the polymerase chain reaction (PCR), has facilitated the use of these probes by greatly increasing the number of copies of target DNA in the sample prior to hybridization. In a conventional assay with immobilized PCR product and labeled oligonucleotide probes, each probe requires a separate hybridization. Here we describe a method by which one can simultaneously screen a sample for all known allelic variants at an amplified locus. In this format, the oligonucleotides are given homopolymer tails with terminal deoxyribonucleotidyltransferase, spotted onto a nylon membrane, and covalently bound by UV irradiation. Due to their long length, the tails are preferentially bound to the nylon, leaving the oligonucleotide probe free to hybridize. The target segment of the DNA sample to be tested is PCR-amplified with biotinylated primers and then hybridized to the membrane containing the immobilized oligonucleotides under stringent conditions. Hybridization is detected nonradioactively by binding of streptavidin-horseradish peroxidase to the biotinylated DNA, followed by a simple colorimetric reaction. This technique has been applied to *HLA-DQA* genotyping (six types) and to the detection of Mediterranean β-thalassemia mutations (nine alleles).

Differential hybridization with sequence-specific oligonucleotide probes has become a widely used technique for the detection of genetic mutations and polymorphisms (1–5). When hybridized under the appropriate conditions, these synthetic DNA probes (usually 15–20 bases in length) will anneal to their complementary target sequences in the sample DNA only if they are perfectly matched. In most cases, the destabilizing effect of a single base-pair mismatch is sufficient to prevent the formation of a stable probe–target duplex (6). With an appropriate selection of oligonucleotide probes, the relevant genetic content of a DNA sample can be completely described.

This very powerful method of DNA analysis has been greatly simplified by the *in vitro* DNA-amplification technique, the polymerase chain reaction (PCR) (7–9). The PCR can selectively increase the number of copies of a particular DNA segment in a sample by many orders of magnitude. As a result of this 10^6- to 10^8-fold amplification, more convenient assays and nonradioactive detection methods have become possible (10–12). These PCR-based assays are usually done by amplifying the target segment in the sample to be tested, fixing the amplified DNA onto a series of nylon membranes, and hybridizing each membrane with one of the labeled oligonucleotide probes under stringent hybridization conditions. However, each probe must still be individually hybridized to the amplified DNA and the process can easily become difficult in a system where many different mutations or polymorphisms occur.

One approach to address this procedural difficulty is to "reverse" the DNAs: attach the oligonucleotides to the nylon support and hybridize the amplified sample to the membrane. Thus, in a single hybridization reaction, an entire series of sequences could be analyzed simultaneously. The strategy we adopted was to immobilize the oligonucleotides onto nylon filters by ultraviolet fixation. Exposure to UV light activates thymine bases in DNA, which then covalently couple to the primary amines present in nylon (13). It seemed unlikely, however, that short oligonucleotides could be directly attached to nylon in this manner and still retain their ability to discriminate at the level of a single base-pair mismatch. Consequently, the addition of a long deoxyribothymidine homopolymer tail, poly(dT), to the 3′ end of the oligonucleotide appeared promising for several reasons. First, the poly(dT) tail would be a larger target for UV crosslinking and should preferentially react with the nylon. Second, dTTP is very readily incorporated onto the 3′ ends of oligonucleotides by terminal deoxyribonucleotidyltransferase and would permit the synthesis of very long tails (14). (Deoxyribothymidine would also be the most efficiently incorporated base if a purely synthetic route were chosen.) Third, Collins and Hunsaker (15) had shown that the presence of a poly(dA) homopolymer tail, used to introduce multiple ^{35}S labels, did not affect the function of sequence-specific oligonucleotide probes.

We have used this technique to attach oligonucleotide probes specific for the six major *HLA-DQA* DNA types (16) and the eight most common Mediterranean β-thalassemia mutations (4) to nylon filters. The target segment of the DNA sample to be tested (either *HLA-DQA* or β-globin) was amplified by PCR with biotin-labeled primers to introduce a nonradioactive tag. Hybridization of the amplified product to the immobilized oligonucleotides and binding of streptavidin-horseradish peroxidase conjugate to the biotinylated primers were performed simultaneously. Detection was accomplished by a simple colorimetric reaction involving the enzymatic oxidation of a colorless chromogen that yielded a red color wherever hybridization occurred.

MATERIALS AND METHODS

Tailing of Oligonucleotides. Oligonucleotides were synthesized on a DNA synthesizer (model 8700, Biosearch) with β-cyanoethyl *N,N*-diisopropylphosphoramidite nucleosides (American Bionetics, Hayward, CA) by using protocols provided by the manufacturer. Oligonucleotide (200 pmol) was tailed in 100 μl of 100 mM potassium cacodylate/25 mM Tris·HCl/1 mM $CoCl_2$/0.2 mM dithiothreitol, pH 7.6 (17), with 5–160 nmol deoxyribonucleoside triphosphate (dTTP or

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: PCR, polymerase chain reaction.

Chee

D–0.1 M KCl, Tat-SF/pp140 was eluted with increasing salt concentrations and was detected mostly in 0.2 to 0.4 M KCl fractions. These fractions were pooled, dialyzed against buffer D–0.1 M KCl, and loaded onto a glutathione Sepharose (Pharmacia) column containing GST-Tat fusion proteins. After the column was washed with buffer D–0.4 M KCl, Tat-SF/pp140 was eluted from the column with buffer D containing 1.4 M KCl. The estimated overall purification after these steps was ~3000-fold. In the experiment shown in Fig. 3, the 0.2 to 0.4 M KCl heparin Sepharose fraction containing Tat-SF activity was subjected to fractionation through an Affi-Gel 10 matrix column (Bio-Rad) containing immobilized Tat. Tat-SF activity was eluted from the column with increasing salt concentrations. The 0.6 M KCl fraction was analyzed as described in Fig. 3.

10. T. O'Brien, S. Hardin, A. Greenleaf, J. T. Lis, *Nature* 370, 75 (1994); M. E. Dahmus, *Biochim. Biophys. Acta* 1261, 171 (1995).
11. A. P. Rice and F. Carlotti, *J. Virol.* 64, 1864 (1990).
12. The Tat-SF/pp140 fraction eluted from the GST-Tat column was subjected to SDS-polyacrylamide gel electrophoresis (PAGE), and the pp140 polypeptide was blotted onto a nitrocellulose membrane. Approximately 15 μg of pp140 were recovered from the membrane and subjected to digestion with lys-C. Six major peptides were obtained and microsequenced. One of the peptides (KMNACETATGNAREERIDE) was contained in the sequence of EST60354 in the Washington University–Merck EST database. An Xho I–Eco RI fragment corresponding to the COOH-terminus of the Tat-SF1 gene and its 3′ untranslated region was labeled and used as a probe to screen a λZipLox (Gibco BRL) cDNA library prepared from human HL60 cells. Complementary DNAs were recovered from seven independent plaques in the autonomously replicating plasmid pZL1 as instructed by the manufacturer (Gibco BRL). The largest cDNA clone containing the full-length Tat-SF1 gene was named pZL-Tat-SF1-4b and was sequenced by dideoxy-DNA sequencing with T7 DNA polymerase.
13. D. R. Marshak and D. Carroll, *Methods Enzymol.* 200, 134 (1991).
14. D. J. Kenan, C. C. Query, J. D. Keene, *Trends Biochem. Sci.* 16, 214 (1991).
15. O. Delattre *et al.*, *Nature* 359, 162 (1992); P. H. Sorensen *et al.*, *Nature Genet.* 6, 146 (1994).
16. A. Crozat, P. Aman, N. Mandahl, D. Ron, *Nature* 363, 640 (1993); T. H. Rabbitts, A. Forster, R. Larson, P. Nathan, *Nature Genet.* 4, 175 (1993).
17. M. Ludanyi, *Diagn. Mol. Pathol.* 4, 162 (1995); T. H. Rabbitts, *Nature* 372, 143 (1994).
18. S. E. Harper, Y. Qiu, P. A. Sharp, *Proc. Natl. Acad. Sci. U.S.A.* 93, 8538 (1996).
19. J. W. Lillie and M. R. Green, *Nature* 338, 39 (1989).
20. H. Kato *et al.*, *Genes Dev.* 6, 655 (1992); R. A. Mardinly and P. A. Sharp, *EMBO J.* 10, 4189 (1991).
21. M. G. Izban and D. S. Luse, *Genes Dev.* 6, 1342 (1992); D. Wang and D. K. Hawley, *Proc. Natl. Acad. Sci. U.S.A.* 90, 843 (1993).
22. E. Bengal, O. Flores, A. Krauskopf, D. Reinberg, Y. Aloni, *Mol. Cell. Biol.* 11, 1195 (1991); J. Greenblatt, J. R. Nodwell, S. W. Mason, *Nature* 364, 401 (1993).
23. C. H. Herrmann and A. P. Rice, *J. Virol.* 69, 1612 (1995).
24. N. A. McMillan *et al.*, *Virology* 218, 413 (1995).
25. W. A. May *et al.*, *Mol. Cell. Biol.* 13, 7393 (1993); H. Zinszner, R. Albalat, D. Ron, *Genes Dev.* 8, 2513 (1994); D. D. Prasad, M. Ouchida, L. Lee, V. N. Rao, E. S. Reddy, *Oncogene* 9, 3717 (1994).
26. P. J. Mitchell and R. Tjian, *Science* 245, 371 (1989).
27. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* 215, 403 (1990).
28. M. A. Trulli *et al.*, *DNA* 4, 339 (1985).
29. H. E. Gendelman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 83, 9759 (1986).
30. L. S. Tiley, P. H. Brown, B. R. Cullen, *Virology* 178, 560 (1990).
31. J. R. Neurmann, C. A. Morency, K. O. Russian, *BioTechniques* 6, 444 (1987).
32. We are grateful to E. Papinsky and Biogen for providing pure HIV Tat protein and Tat mutant TatAC; to J. Borrow (Massachusetts Institute of Technology (MIT) Center for Cancer Research) for human cDNA libraries; and to R. Cook (MIT Biopolymers Laboratory) for peptide sequencing. We thank K. Luo, J. Burrow, and H. Kawasaki for valuable advice and discussions, and B. Ellencowe, K. Cepek, G. Jones, K. Luo, and C. Query for helpful comments on the manuscript. We also thank M. Sistare for secretarial support. Supported by grants from the National Institutes of Health (GM34277 and AI32488) to P.A.S., and partially supported by a National Cancer Institute Center core grant (CA14051). Q.Z. was supported by a postdoctoral fellowship of The Jane Coffin Childs Memorial Fund for Medical Research.

19 June 1996; accepted 23 August 1996

Accessing Genetic Information with High-Density DNA Arrays

Mark Chee, Robert Yang, Earl Hubbell, Anthony Berno, Xiaohua C. Huang, David Stern, Jim Winkler, David J. Lockhart, Macdonald S. Morris, Stephen P. A. Fodor

Rapid access to genetic information is central to the revolution taking place in molecular genetics. The simultaneous analysis of the entire human mitochondrial genome is described here. DNA arrays containing up to 135,000 probes complementary to the 16.6-kilobase human mitochondrial genome were generated by light-directed chemical synthesis. A two-color labeling scheme was developed that allows simultaneous comparison of a polymorphic target to a reference DNA or RNA. Complete hybridization patterns were revealed in a matter of minutes. Sequence polymorphisms were detected with single-base resolution and unprecedented efficiency. The methods described are generic and can be used to address a variety of questions in molecular genetics including gene expression, genetic linkage, and genetic variability.

A central theme in modern genetics is the relation between genetic variability and phenotype. To understand genetic variation and its consequences on biological function, an enormous effort in comparative sequence analysis will need to be carried out. Conventional nucleic acid sequencing technologies make use of analytical separation techniques to resolve sequence at the single nucleotide level (1, 2). However, the effort required increases linearly with the amount of sequence. In contrast, biological systems read, store, and modify genetic information by molecular recognition (3). Because each DNA strand carries with it the capacity to recognize a uniquely complementary sequence through base pairing, the process of recognition, or hybridization, is highly parallel, as every nucleotide in a large sequence can in principle be queried at the same time. Thus, hybridization can be used to efficiently analyze large amounts of nucleotide sequence. In one proposal, sequences are analyzed by hybridization to a set of oligonucleotides representing all possible subsequences (4). A second approach, used here, is hybridization to an array of oligonucleotide probes designed to match specific sequences. In this way the most informative subset of probes is used. Implementation of these concepts relies on recently developed combinatorial technologies to generate any ordered array of a large number of oligonucleotide probes (5).

Affymetrix, 3380 Central Expressway, Santa Clara, CA 95051, USA.

The fundamentals of light-directed oligonucleotide array synthesis have been described (5, 6). Any probe can be synthesized at any discrete, specified location in the array, and any set of probes composed of the four nucleotides can be synthesized in a maximum of $4N$ cycles, where N is the length of the longest probe in the array. For example, the entire set of $\sim 10^{12}$ 20-nucleotide oligomer probes, or any desired subset, can be synthesized in only 80 coupling cycles. The number of different probes that can be synthesized is limited only by the physical size of the array and the achievable lithographic resolution (7).

An array consisting of oligonucleotides complementary to subsequences of a target sequence can be used to determine the identity of a target sequence, measure its amount, and detect differences between the target and a reference sequence. Many different arrays can be designed for these purposes. One such design, termed a 4L tiled array, is depicted in Fig. 1A. In each set of four probes, the perfect complement will hybridize more strongly than mismatched probes. By this approach, a nucleic acid target of length L can be scanned for mutations with a tiled array containing 4L probes. For example, to query the 16,569 base pairs (bp) of human mitochondrial DNA (mtDNA), only 66,276 probes of the possible $\sim 10^9$ 15-nucleotide oligomers need to be used.

The use of a tiled array of probes to read a target sequence is illustrated in Fig. 1C. A tiled array of 15-nucleotide oligomers varied

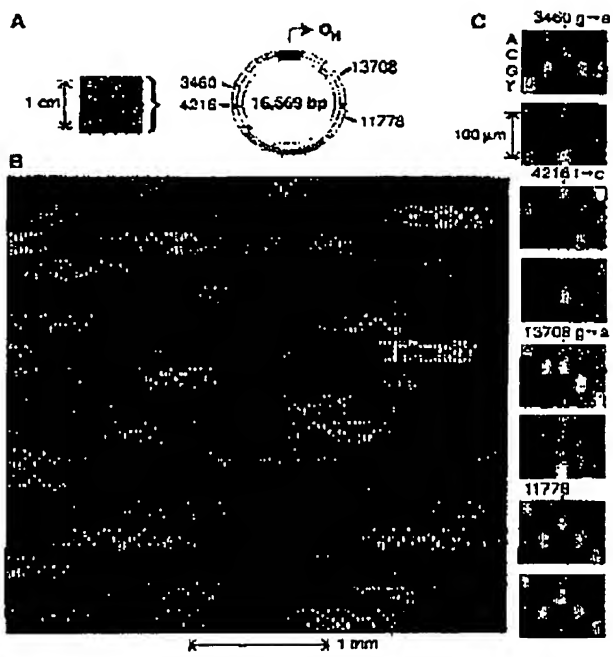


Fig. 3. Human mitochondrial genome on a chip. (A) An image of the array hybridized to 16.6 kb of mitochondrial target RNA (L strand). The 16,569-bp map of the genome is shown, and the H strand origin of replication (O_H), located in the control region, is indicated. (B) A portion of the hybridization pattern magnified. In each column there are five probes: A, C, G, T, and Δ, from top to bottom. The Δ probe has a single-base deletion instead of a substitution and hence is 24 instead of 25 bases in length. The scale is indicated by the bar beneath the image. Although there is considerable sequence-dependent intensity variation, most of the array can be read directly. The image was collected at a resolution of ~100 pixels per probe cell. (C) The ability of the array to detect and read single-base differences in a 16.6-kb sample is illustrated. Two different target sequences were hybridized in parallel to different chips. The hybridization patterns are compared for four different positions in the sequence. Only the $P^{26,13}$ probes are shown. The top panel of each pair shows the hybridization of the mt3 target, which matches the chip P^0 sequence at these positions. The lower panel shows the pattern generated by a sample from a patient with Leber's hereditary optic neuropathy (LHON). Three known pathogenic mutations, LHON3460, LHON4216, and LHON13708, are clearly detected. For comparison, the fourth panel in the set shows a region around position 11,778 that is identical in both samples.

provide the foundation for a powerful generic analysis technology. The method can be used to characterize the spectrum of sequence variation in a population and can be applied to the analysis of many genes in parallel. In the case of human mtDNA, we simultaneously analyzed the control region, 13 protein coding genes, 22 tRNA genes, and 2 ribosomal RNA genes. The methods described here can be applied to other research areas in molecular genetics. For example, the ability to identify and sequence polymorphisms provides a basis for genetic mapping. The specificity of oligonucleotide hybridization and the scalability of the method suggests the possibility of a dedicated array that could be used to generate a high-resolution genetic map of an entire genome in a single experiment. Likewise, the concepts and techniques described here have been used to develop approaches for mRNA identification and the large-scale, parallel measurement of expression levels (24). Thus, the sequence of a gene, its spectrum of change in the population, its chromosomal location, and its dynamics of expression (all essential to a full understanding of function) can be determined with high-density probe arrays. The challenge now is to synthesize and read probe arrays at even higher density. For example, a 2 cm by 2 cm array, synthesized with probes occupying 1-μm synthesis sites in a 4L tiling, could query the entire coding content of the human genome, estimated at 100,000 genes.

REFERENCES AND NOTES

1. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* 74, 5463 (1977).
2. A. M. Maxam and W. Gilbert, *ibid.*, p. 560.
3. J. D. Watson and F. H. C. Crick, *Nature* 171, 737 (1953).
4. W. Bains and G. C. Smith, *J. Theor. Biol.* 135, 303 (1988); Y. P. Lysov *et al.*, *Dokl. Akad. Nank. SSSR* 303, 1508 (1988); R. Drmanac, I. Labat, I. Brukner, R. Crkvenjakov, *Genomics* 4, 114 (1989); E. Southern, U. Maskos, R. Elder, *ibid.* 13, 1008 (1992); see also R. B. Wallace *et al.*, *Nucleic Acids Res.* 6, 3543 (1979).
5. S. P. A. Fodor *et al.*, *Science* 251, 767 (1991).
6. A. C. Pease *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 91, 5022 (1994).
7. In the present format, we can routinely achieve a density of 409,600 synthesis sites in a 1.28 cm by 1.28 cm array. Each 20 μm by 20 μm site contains ~4 × 10^6 functional copies of a specific probe, which corresponds to a mean distance of about 100 Å between probes (M. O. Trulson, D. Stern, R. P. Rava, unpublished results).
8. S. Anderson *et al.*, *Nature* 290, 457 (1981).
9. The control region of mtDNA is characterized by high amounts of sequence polymorphism concentrated in two hypervariable regions [B. D. Greenberg, J. E. Newbold, A. Sugino, *Gene* 21, 33 (1983)]; C. F. Aquadro and B. D. Greenberg, *Genetics* 103, 287 (1983)].
10. R. L. Cann, W. M. Brown, A. C. Wilson, *Genetics* 106, 479 (1984).
11. The mt1 and mt2 sequences were cloned from amplified genomic DNA extracted from hair roots [P. Gill, A. J. Jeffreys, D. J. Werrett, *Nature* 318, 577 (1985); R. K. Saiki *et al.*, *Science* 239, 487 (1988)]. The clones were sequenced conventionally (7). Cloning was performed only to provide a set of pure reference samples of known sequence. For templates for fluorescent labeling, DNA was reamplified from the clones with primers bearing bacteriophage T3 and T7 RNA polymerase promoter sequences (bold: mtDNA sequences uppercase): L16935-T3, 5′-**ctcaggaattaaccctcactaaagg**AAACCTTTTTCC-AAGGA and H687-T7, 5′-**taatacgactcactataggga**-gAGGCTAGGACCAAACCTATT.
12. Labeled RNAs from the two complementary mtDNA strands (designated L and H (8)) were transcribed in separate reactions from a promoter-tagged polymerase chain reaction (PCR) product. Each 10-μl reaction contained 1.5 mM each of the triphosphate nucleotides ATP, CTP, GTP, and UTP; 0.24 mM fluorescein-12-CTP (Du Pont); 0.24 mM fluorescein-12-UTP (Boehringer Mannheim); ~1 to 5 nM (1.5 μl) crude unpurified 1.3-kb PCR product; and T3 or T7 RNA polymerase (1 U/μl) (Promega) in a reaction buffer supplied with the enzyme. The reaction was carried out at 37°C for 1 to 2 hours. RNA was fragmented to an average size of <100 nucleotides by adjusting the solution to 30 mM $MgCl_2$ by the addition of 1 M $MgCl_2$ and heating at 94°C for 40 min. Fragmentation improved the uniformity and specificity of hybridization (M. Chee *et al.*, data not shown). The extent of fragmentation is dependent on the magnesium ion concentration [J. W. Huff, K. S. Sastry, M. P. Gordon, W. E. C. Wacker, *Biochemistry* 3, 501 (1964); J. J. Butzow and G. L. Eichorn, *Biopolymers* 3, 95 (1965)]. Good hybridization results have been obtained with both DNA and RNA targets prepared with a variety of labeling schemes, including incorporation of fluorescent and biotinylated deoxynucleoside triphosphates by DNA polymerases, incorporation of dye-labeled primers during PCR, ligation of labeled oligonucleotides to fragmented RNA, and direct labeling by photo-cross-linking a psoralen derivative of biotin directly to fragmented nucleic acids (L. Wodicka, personal communication).
13. For two-color detection experiments, the reference and unknown samples were labeled with biotin and fluorescein, respectively, in separate transcription reactions. Reactions were carried out as described (12) except that each contained 1.25 mM of ATP, CTP, GTP, and UTP and 0.5 mM fluorescein-12-... heim]. The two reactions were mixed in the ratio 1:5 (w/w) biotin:fluorescein and fragmented (12). Targets were diluted to a final concentration of ~100 to 1000 pM in 3M TMACl [W. B. Melchior Jr. and P. H. von Hippel, *Proc. Natl. Acad. Sci. U.S.A.* 70, 298 (1973)], 10 mM tris-HCl, pH 8.0, 1 mM EDTA, 0.005% Triton X-100, and 0.2 nM control oligonucleotide labeled at the 5′ end with fluorescein (5′-CTGAACGGTAG-CATCTTGAC). Samples were denatured at 95°C for 5 min, chilled on ice for 5 min, and equilibrated to 37°C. A volume of 180 μl of hybridization solution was then added to the flow cell [R. Lipshutz *et al.*, *Biotechniques* 19, 442 (1995)] and the chip incubated at 37°C for 3 hours with rotation at 60 rpm. The chip was washed six times at room temperature with 6× SSPE (0.9 M NaCl, 60 mM NaH_2PO_4, 6 mM EDTA, pH 7.4), 0.005% Triton X-100. Phycoerythrin-conjugated streptavidin (2 μg/ml in 6× SSPE, 0.005% Triton X-100) was added and incubation continued at room temperature for 5 min. The chip was washed again

The Future of Genetic Studies of Complex Human Diseases

Neil Risch and Kathleen Merikangas

Geneticists have made substantial progress in identifying the genetic basis of many human diseases, at least those with conspicuous determinants. These successes include Huntington's disease, Alzheimer's disease, and some forms of breast cancer. However, the detection of genetic factors for complex diseases—such as schizophrenia, bipolar disorder, and diabetes—has been far more complicated. There have been numerous reports of genes or loci that might underlie these disorders, but few of these findings have been replicated. The modest nature of the gene effects for these disorders likely explains the contradictory and inconclusive claims about their identification. Despite the small effects of such genes, the magnitude of their attributable risk (the proportion of people affected due to them) may be large because they are quite frequent in the population, making them of public health significance.

Has the genetic study of complex disorders reached its limits? The persistent lack of replicability of these reports of linkage between various loci and complex diseases might imply that it has. We argue below that the method that has been used successfully (linkage analysis) to find major genes has limited power to detect genes of modest effect, but that a different approach (association studies) that utilizes candidate genes has far greater power, even if one needs to test every gene in the genome. Thus, the future of the genetics of complex diseases is likely to require large-scale testing by association analysis.

How large does a gene effect need to be in order to be detectable by linkage analysis? We consider the following model: Suppose a disease susceptibility locus has two alleles A and a, with population frequencies p and $q = 1 - p$, respectively. There are three genotypes: AA, Aa, and aa. We define genotypic relative risks (GRR, the increased chance that an individual with a particular genotype has the disease) as follows: Let the risk for individuals of genotype Aa be γ times greater than the risk for individuals with genotype aa, a GRR of γ. We assume a multiplicative relation for two A alleles, so that the GRR for genotype AA is γ^2. The method of linkage analysis we have chosen for this argument is a popular current paradigm in which pairs of siblings, both with the disease, are examined for sharing of alleles at multiple sites in the genome defined by genetic markers. The more often the affected siblings share the same allele at a particular site, the more likely the site is close to the disease gene. Using the formulas in (*1*), we calculate the expected proportion Y of alleles shared by a pair of affected siblings for the best possible case—that is, a closely linked marker locus (recombination fraction $\theta = 0$) that is fully informative (heterozygosity = 1) (*2*)—as

$$Y = \frac{1 + w}{2 + w} \text{ where } w = \frac{pq(\gamma - 1)^2}{(p\gamma + q)^2}$$

If there is no linkage of a marker at a particular site to the disease, the siblings would be expected to share alleles 50% of the time; that is, Y would equal 0.5. Values of Y for various values of p and γ are given in the third column of the table. For an allele of moderate frequency (p is 0.1 to 0.5) that confers a GRR (γ) of fourfold or greater, there is a detectable deviation of Y from the null value of 0.5. On the other hand, for an allele conferring a GRR of 2 or less, the expected marker-sharing only marginally exceeds 50%, for any allele frequency (p). Thus, it is clear that the use of linkage analysis for loci conferring GRR of about 2 or less will never allow identification because the number of families required (more than ~2500) is not practically achievable.

Although tests of linkage for genes of modest effect are of low power, as shown by the above example, direct tests of association with a disease locus itself can still be quite strong. To illustrate this point, we use the transmission/disequilibrium test of Spielman *et al.* (*3*). In this test, transmission of a particular allele at a locus from heterozygous parents to their affected offspring is examined. Under Mendelian inheritance, all alleles should have a 50% chance of being transmitted to the next generation. In contrast, if one of the alleles is associated with disease risk, it will be transmitted more often than 50% of the time.

For this approach, we do not need families with multiple affected siblings, but can focus just on single affected individuals and their parents. For the same model given above, we can calculate the proportion of heterozygous parents as $pq(\gamma + 1)/(p\gamma + q)(4)$. Similarly, the probability for a heterozygous parent to transmit the high risk A allele is just $\gamma/(1 + \gamma)$. Association tests can also be performed for pairs of affected siblings. When the locus is associated with disease, the transmission excess over 50% is the same as for single offspring, but the probability of parental heterozygosity is increased at low values of p; for higher values of p, the probability of parental heterozygosity is decreased. The formula for parental heterozygosity for an affected pair of siblings for the same genetic model as used in the first example is

$$h = \frac{pq(\gamma + 1)^2}{2(p\gamma + q)^2 + pq(\gamma - 1)^2}$$

		Linkage			Association			
					Singletons		Sib pairs	
Genotypic risk ratio (γ)	Frequency of disease allele A (p)	Probability of allele sharing (Y)	No. of families required (N)	Probability of transmitting disease allele A $P(tr\text{-}A)$	Proportion of heterozygous parents (Het)	(N)	(Het)	(N)
4.0	0.01	0.520	4260	0.800	0.048	1098	0.112	235
	0.10	0.597	185	0.800	0.346	150	0.537	48
	0.50	0.578	297	0.800	0.500	103	0.424	61
	0.80	0.529	2013	0.800	0.235	222	0.163	161
2.0	0.01	0.502	296,710	0.667	0.029	5823	0.043	1970
	0.10	0.518	5382	0.667	0.245	605	0.323	264
	0.50	0.526	2498	0.667	0.500	340	0.474	180
	0.80	0.512	11,917	0.667	0.267	640	0.217	394
1.5	0.01	0.501	4,620,807	0.600	0.025	19,220	0.031	7776
	0.10	0.505	67,816	0.600	0.197	2218	0.253	941
	0.50	0.510	17,997	0.600	0.500	949	0.490	484
	0.80	0.505	67,816	0.600	0.286	1663	0.253	941

Comparison of linkage and association studies. Number of families needed for identification of a disease gene.

N. Risch is in the Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305–5120, USA. E-mail: risch@lishmed.stanford.edu. K. Merikangas is in the Departments of Epidemiology and Psychiatry, Unit, Yale University School of Medicine, New Haven, CT 06510, USA. E-mail: kath@zeus.psych.yale.edu

Risch

On the right side of the table, we present the proportion of heterozygous parents (Het) and the probability of transmission of the A allele from a heterozygous parent to an affected child [$P(tr\text{-}A)$] for the same values of GRR as considered above for the example of linkage analysis. The deviation from the null hypothesis of 50% transmission from heterozygous parents is substantially greater than the excess allele sharing that is found by linkage analysis in sibling pairs. This disparity between the methods is particularly true for lower values of γ (that is, with lower relative risk). For example, for $\gamma = 1.5$, allele sharing is at most 51%, while the A allele is transmitted 60% of the time from heterozygous parents.

In this respect then, association studies seem to be of greater power than linkage studies. But of course, the limitation of association studies is that the actual gene or genes involved in the disease must be tentatively identified before the test can be performed. In fact, the actual polymorphism within the gene (or at least a polymorphism in strong disequilibrium) must be available. However, we show that this requirement is only daunting because of limitations imposed by current technological capabilities, not because sufficient families with the disease are not available or the statistical power is inadequate (5). For example, imagine the time when all human genes (say 100,000 in total) have been found and that simple, diallelic polymorphisms in these genes have been identified. Assume that five such diallelic polymorphisms have been identified within each gene, so that a total of $10 \times 10^5 = 10^6$ alleles need to be tested. The statistical problem is that the large number of tests that need to be made leads to an inflation of the type 1 error probability. For a linkage test with pairs of affected siblings, we use a lod score (logarithm of the odds ratio for linkage) criterion of 3.0, which asymptotically corresponds to a type 1 error probability α of about 10^{-4}. In a linkage genome screen with 500 markers, this significance level gives a probability greater than 95% of no false positives. The equivalent false positive rate for 1,000,000 independent association tests can be obtained with a significance level $\alpha = 5 \times 10^{-8}$.

We illustrate the power of linkage versus association tests at different significance levels by determining the sample size N (number of families) necessary to obtain 80% power (the probability of rejecting the null hypothesis when it is false) (6) (see table). With a linkage approach and a disease gene with a GRR of 4 or greater, the number of affected sibling pairs necessary to detect linkage is realistic (185 or 297), provided the allele frequency p is between 5 and 75%. For a gene with a GRR of 2 or less, however, the sample sizes are generally beyond reach (well over 2000), precluding their identification by this approach. In contrast, the required sample size for the association test, even allowing for the smaller significance level, is vastly less than for linkage, especially for affected sibling pair families when the value of p is small. Even for a GRR of 1.5, the sample sizes are generally less than 1000, well within reason.

Thus, the primary limitation of genome-wide association tests is not a statistical one but a technological one. A large number of genes (up to 100,000) and polymorphisms (preferentially ones that create alterations in derived proteins or their expression) must first be identified, and an extremely large number of such polymorphisms will need to be tested. Although testing such a large number of polymorphisms on several hundred, or even a thousand families, might currently seem implausible in scope, more efficient methods of screening a large number of polymorphisms (for example, sample pooling) may be possible. Furthermore, the number of tests we have used as the basis for our calculations (1,000,000) is likely to be far larger than necessary if one allows for linkage disequilibrium, which could substantially reduce the required number of markers and families needed for initial screening.

Some of the important loci for complex diseases will undoubtedly be found by linkage analysis. However, the limitations to detecting many of the remaining genes by linkage studies can be overcome; numerous genetic effects too weak to identify by linkage can be detected by genomic association studies. Fortunately, the samples currently collected for linkage studies (for example, affected pairs of siblings and their parents) can also be used for such association studies. Thus, investigators should preserve their samples for future large-scale testing.

The human genome project can have more than one reward. In addition to sequencing the entire human genome, it can lead to identification of polymorphisms for all the genes in the human genome and the diseases to which they contribute. It is a charge to the molecular technologists to develop the tools to meet this challenge and provide the information necessary to identify the genetic basis of complex human diseases.

References and Notes

1. N. Risch, *Am. J. Hum. Genet.* 40, 1 (1987); *ibid.* 46, 229 (1990).
2. From the formulas in (1), we have $\lambda_O = 1 + 0.5V_A/K^2$ and $\lambda_S = 1 + (0.5V_A + 0.25V_D)/K^2$, where $K = p^2\gamma^2 + 2pq\gamma + q^2 = (p\gamma + q)^2$, $V_A = 2pq(\gamma - 1)^2\,(p\gamma + q)^2$, and $V_D = p^2q^2(\gamma - 1)^2$. Hence, $\lambda_D = 1 + w$ and $\lambda_S = (1 + 0.5w)^2$, where $w = pq(\gamma - 1)^2$. The proportion of alleles shared is given by $Y = 1 - 0.5z_1 - z_0$, where z_1 and z_0 are the probabilities of the sib pair sharing 1 and 0 disease alleles ibd, respectively. From (1), $z_0 = 0.25/\lambda_S$ and $z_1 = 0.5\lambda_O/\lambda_S$. Thus, after some algebra, $Y = 1 - 0.25(\lambda_O + 1)/\lambda_S = (1 + w)/(2 + w)$.
3. R. Spielman, R. E. McGinnis, W. J. Ewens, *Am. J. Hum. Genet.* 52, 506 (1993).
4. By Bayes' theorem, the probability of a parent of an affected child being heterozygous is given by $P(\text{Het}|\text{Aff child}) = P(\text{Het})P(\text{Aff Child}|\text{Het})/P(\text{All Child}) = 2pq[0.5p(\gamma^2 + \gamma)] \div 0.5q(\gamma + 1)/(p\gamma + q)^2 = pq(\gamma + 1)/(p\gamma + q)$.
5. E. S. Lander and N. J. Schork, *Science* 265, 2037 (1994).
6. Consider a set of M independent, identically distributed random variables B_i of discrete value. Under the null hypothesis H_0, assume $E(B_i) = 0$ and $\text{Var}(B_i) = 1$. Under the alternative hypothesis H_1, let $E(B_i) = \mu$ and $\text{Var}(B_i) = \sigma^2$. For a sample of size M, let $T = \Sigma B_i/\sqrt{M}$. Then under H_0, T also has mean 0 and variance 1, while under H_1, it has mean $\sqrt{M}\mu$ and variance σ^2. We assume that T is approximately normally distributed both under H_0 and H_1. Then the sample size M required to obtain a power of $1 - \beta$ for a significance level α is given by

$$M = (Z_\alpha - \sigma Z_{1-\beta})^2/\mu^2 \quad (1)$$

For each affected sib pair, we score the number of alleles shared ibd from each of $2N$ parents. Define $B_i = 1$ if an allele is shared from the ith parent and $B_i = -1$ if unshared. Under the null hypothesis of no linkage, $P(B_i = 1) = P(B_i = -1) = 0.5$, so $E(B_i) = 0$ and $\text{Var}(B_i) = 1$. For the genetic model described above with genotypic relative risks of γ^2, γ, and 1, allele sharing by affected sibs is independent for the two parents; thus, we can consider sharing of alleles one parent at a time. Thus, for affected sib pairs assuming $\theta = 0$ and no linkage disequilibrium, the formula is

$$N = \frac{(Z_\alpha - \sigma Z_{1-\beta})^2}{2\mu^2}$$

where

$$\mu = 2Y - 1$$

$$\sigma^2 = 4Y(1 - Y)$$

$$Y = \frac{1 + w}{2 + w}$$

$$w = \frac{pq\,(\gamma - 1)^2}{(p\gamma + q)^2}$$

$Z_\alpha = 3.72$ (corresponding to $\alpha = 10^{-4}$), and $Z_{1-\beta} = -0.84$ (corresponding to $1 - \beta = 0.80$). For an association test using the transmission/disequilibrium test, with the disease locus or a nearby locus in complete disequilibrium, the number (N) of families with affected singletons required for 80% power is also calculated from formula 1. For this case, we score the number of transmissions of allele A from heterozygous parents. Let h be the probability a parent is heterozygous under the alternative hypothesis, namely, $h = pq(\gamma + 1)/(p\gamma + q)$. Then define $B_i = h^{-0.5}$ if the parent is heterozygous and allele A is transmitted; $B_i = 0$ if the parent is homozygous; and $B_i = -h^{-0.5}$ if the parent is heterozygous and transmits allele a. Under the null hypothesis, $E(B_i) = 0$ and $\text{Var}(B_i) = 1$. Under the alternative hypothesis, $\mu = E(B_i) = \sqrt{h}(\gamma - 1)/(\gamma + 1)$ and $\sigma^2 = \text{Var}(B_i) = 1 - h(\gamma - 1)^2/(\gamma + 1)^2$. In this case, there are two parents per family and they act independently, so the required number (N) of families is given by half of formula 1 where μ and σ^2 are given above. Here, $Z_\alpha = 5.33$ (corresponding to $\alpha = 5 \times 10^{-8}$). For the same test but with affected sib pairs instead of singletons, the number of families required is given by half of formula 1 (transmissions from two parents to two children) with the same formulas for μ and σ^2 as for singleton families but now using the heterozygote frequency for parents of affected sib pairs. Using the above formulas, we can calculate sample sizes for the three study designs.

27 October 1995; accepted 6 June 1996.

Document for use in Amendment/Responses 9/13/05 and 11/05 for application 10/037, 718 1
Applicants MCGINNIS ET AL

Illustration of an Example Set/Subset N-covering using a CL-F map: Subsets of bi-allelic covering markers N-cover an entire, large rectangular CL-F region bounded by a chromosome or chromosomal subregion of interest in the chromosomal location (CL) dimension and bounded by the range 0 to 0.5 in the (least common allele) frequency (F) dimension.

Subsets of bi-allelic covering markers are chosen whereby each of the 10 smaller rectangular segment-subranges designated A, B, C, D or E (on the left and right) contains two or more covering markers that belong to the same subset. These 10 overlapping segment-subranges completely cover the entire, large rectangular CL-F region (arrows pointing to the top boundary and right boundary). Each of these 10 segment-subranges is less than or equal to L_2 in length and equal to 0.15 in width. So each point in the entire large rectangular region is within the two-dimensional (CL-F) distance [L_2, 0.15] of two or more covering markers. **That is, the entire large rectangular region is N-covered to within [L_2, 0.15] by the bi-allelic covering markers, wherein N ≥ 2.**